Mathmatial Foundations of Reinforcement Learning

Lower Bounds

Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

Fall 2023

Outline

Warm-up: basic tools

Lower bounds for multi-arm bandits

Analysis

Motivation

So far, we have seen that for both stochastic bandits and adversarial bandits, the worst-case regret bound R_T scales as (ignoring logarithmic factors)

$$\widetilde{O}(\sqrt{T}).$$

Question

Can we improve the worst-case regret, say to $\widetilde{O}(T^{1/4})$ or $\widetilde{O}(T^{1/3})$?

Motivation

So far, we have seen that for both stochastic bandits and adversarial bandits, the worst-case regret bound R_T scales as (ignoring logarithmic factors)

$$\widetilde{O}(\sqrt{T}).$$

Question

Can we improve the worst-case regret, say to $\widetilde{O}(T^{1/4})$ or $\widetilde{O}(T^{1/3})$?

Two paths:

- Try hard to come up with a better algorithm.

Motivation

So far, we have seen that for both stochastic bandits and adversarial bandits, the worst-case regret bound R_T scales as (ignoring logarithmic factors)

$$\widetilde{O}(\sqrt{T}).$$

Question

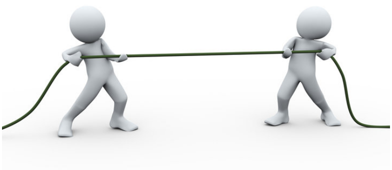
Can we improve the worst-case regret, say to $\widetilde{O}(T^{1/4})$ or $\widetilde{O}(T^{1/3})$?

Two paths:

- Try hard to come up with a better algorithm.
- Develop negative results that show this is impossible. Our plan!

Why studying lower bounds?

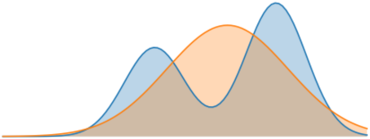
- Lower bounds tell us the minimal price we need to pay.
- Benchmark performance: given an upper bound for certain algorithm, how much room can we improve?
- Matching upper and lower bounds tells us the **fundamental limits**.



Warm-up: basic tools

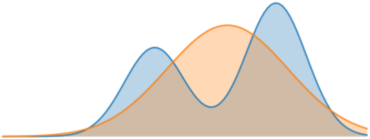
Which distributions do the data come from?

- Consider hypothesis testing.
- Under different hypotheses, we collect data with different distributions
- The (in)ability to distinguish these distributions becomes the key



Which distributions do the data come from?

- Consider hypothesis testing.
- Under different hypotheses, we collect data with different distributions
- The (in)ability to distinguish these distributions becomes the key



Question: how do we measure the distance between distributions?

Distance of distributions

Definition 1 (TV distance)

For two probabilities p, q over Ω, the total variation distance is given by

$$d_{\mathsf{TV}}(p,q) = \sup_{A \subseteq \Omega} \{p(A) - q(A)\} \in [0,1].$$

Distance of distributions

Definition 1 (TV distance)

For two probabilities p, q over Ω, the total variation distance is given by

$$d_{\mathsf{TV}}(p,q) = \sup_{A \subseteq \Omega} \{p(A) - q(A)\} \in [0,1].$$

Definition 2 (KL divergence)

For two probabilities p, q over Ω, the Kullback-Leibler (KL) divergence is given by

$$\mathsf{KL}(p \| q) = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)}.$$

Examples of KL divergence

- **Bernoulli distributions:**

$$\begin{aligned}\mathsf{KL}\big(\mathsf{Bern}(\tfrac{1+\epsilon}{2})\|\mathsf{Bern}(\tfrac{1}{2})\big) &= \frac{1+\epsilon}{2}\log\big(1+\epsilon\big) + \frac{1-\epsilon}{2}\log\big(1-\epsilon\big) \\ &\le 2\epsilon^2,\end{aligned}$$

  where the inequality follows from $\log(1+x) \le x$. Note that

$$\mathsf{KL}\big(\mathsf{Bern}(\tfrac{1}{2})\|\mathsf{Bern}(\tfrac{1+\epsilon}{2})\big) \neq \mathsf{KL}\big(\mathsf{Bern}(\tfrac{1+\epsilon}{2})\|\mathsf{Bern}(\tfrac{1}{2})\big)!$$

- **Gaussian distributions:**

$$\mathsf{KL}\big(\mathcal{N}(\mu_1,\sigma^2)\|\mathcal{N}(\mu_2,\sigma^2)\big) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}$$

Pinsker's inequality

Pinsker's inequality

For two probability distributions p, q over Ω, it holds that

$$2d_{\mathsf{TV}}(p,q)^2 \leq \mathsf{KL}(p\|q).$$

- By definition, for any event $A \subseteq \Omega$,

$$p(A) - q(A) \leq \sqrt{\frac{1}{2}\mathsf{KL}(p\|q)}.$$

- Due to asymmetry of KL divergence, we have:

$$2d_{\mathsf{TV}}(p,q)^2 \leq \min\left\{\mathsf{KL}(p\|q), \mathsf{KL}(p\|q)\right\}.$$

- A very useful tool!

Toy example: binary hypothesis testing

Suppose we observe a sequence of coin flips

$$A_t \overset{\text{i.i.d.}}{\sim} \mathsf{Ber}(\mu), \quad t = 1, \ldots, T.$$

Consider two hypotheses for μ:

$$\mathcal{H}_0: \quad \mu = \frac{1}{2}, \qquad \mathcal{H}_1: \quad \mu = \frac{1+\epsilon}{2}.$$

We want to determine which hypothesis is true. Is the coin fair?

Question

How many samples do we need to collect in order to do so reliably?

Step 1: KL divergence of the data

Denote the data distribution under two hypotheses respectively as

$$\mathbb{P}_0 := \mathbb{P}\left(A_1, A_2 \ldots, A_T | \mathcal{H}_0\right)$$
$$\mathbb{P}_1 := \mathbb{P}\left(A_1, A_2 \ldots, A_T | \mathcal{H}_1\right).$$

Then, it is easy to see

$$\begin{aligned}\mathsf{KL}(\mathbb{P}_1 \| \mathbb{P}_0) &= \sum_{i=1}^{T} \mathsf{KL}\left(\mathbb{P}(A_i|\mathcal{H}_1) \| \mathbb{P}(A_i|\mathcal{H}_0)\right) \\ &= T \cdot \mathsf{KL}\big(\mathsf{Bern}(\tfrac{1+\epsilon}{2}) \| \mathsf{Bern}(\tfrac{1}{2})\big) \\ &\leq 2T\epsilon^2.\end{aligned}$$

The KL divergence scales linear in T and quadratically in ϵ.

Step 2: determine the goal

Question: what do we mean by solving the problem "reliably"?

Answer: Maybe getting a correct answer with non-trivial probability, e.g. for some small probability of error δ,

$$\mathbb{P}\left(\text{learner outputs fair}|\mathcal{H}_0\right) \geq 1 - \delta/2.$$
$$\mathbb{P}\left(\text{learner outputs unfair}|\mathcal{H}_1\right) \geq 1 - \delta/2.$$

Let us call the event $A = \{\text{learner outputs fair}\}$, then the above leads to

$$\mathbb{P}_0(A) \geq 1 - \delta/2, \qquad \mathbb{P}_1(A) \leq \delta/2$$

Step 2: determine the goal

Question: what do we mean by solving the problem "reliably"?

Answer: Maybe getting a correct answer with non-trivial probability, e.g. for some small probability of error δ,

$$\mathbb{P}\left(\text{learner outputs fair}|\mathcal{H}_0\right) \geq 1 - \delta/2.$$
$$\mathbb{P}\left(\text{learner outputs unfair}|\mathcal{H}_1\right) \geq 1 - \delta/2.$$

Let us call the event $A = \{\text{learner outputs fair}\}$, then the above leads to

$$\mathbb{P}_0(A) \geq 1 - \delta/2, \qquad \mathbb{P}_1(A) \leq \delta/2$$

$$\Longrightarrow \qquad \mathbb{P}_0(A) - \mathbb{P}_1(A) \geq 1 - \delta.$$

Step 3: applying Pinsker

By Pinsker's inequality, we know

$$2(\mathbb{P}_0(A) - \mathbb{P}_1(A))^2 \le \mathsf{KL}(\mathbb{P}_1 \| \mathbb{P}_2) \le 2T\epsilon^2.$$

Hence,

$$\begin{aligned} T &\ge \frac{(\mathbb{P}_0(A) - \mathbb{P}_1(A))^2}{\epsilon^2} \\ &\ge \frac{(1-\delta)^2}{\epsilon^2}. \end{aligned}$$

The sample size T needs to be at least

$$T \gtrsim \frac{1}{\epsilon^2}!$$

Examine the upper bound

The scaling $T \gtrsim \frac{1}{\epsilon^2}$ turns out to be sufficient too!

By Hoeffding's inequality, we know

$$\left| \frac{1}{n} \sum_{t=1}^{T} A_t - \mu \right| \leq \sqrt{\frac{\log(2/\delta)}{2T}} \qquad \text{with prob. } 1 - \delta.$$

By setting $\sqrt{\frac{\log(2/\delta)}{2T}} \leq \frac{\epsilon}{4}$, or equivalently, $T \geq \frac{8 \log(2/\delta)}{\epsilon^2}$, we guarantee

$$\left| \frac{1}{n} \sum_{t=1}^{T} A_t - \mu \right| \leq \frac{\epsilon}{4} \qquad \text{with prob. } 1 - \delta.$$

The learner compares the sample mean $\frac{1}{n} \sum_{t=1}^{T} A_t$ with $\frac{1}{2} + \frac{\epsilon}{4}$.

Lower bounds for multi-arm bandits

Worst-case lower bound

For simplicity, we will assume all arms have a Gaussian reward distribution $\mathcal{N}(\mu_i, 1)$ for $i \in [n]$.

Theorem 3 (minimax lower bound)

Let $n > 1$ and $T \geq n - 1$. Then, for any algorithm π, there exists a mean vector $\mu = [\mu_i]_{1 \leq i \leq n} \in [0, 1]^n$ such that

$$R_T \geq \frac{1}{27}\sqrt{(n-1)T}.$$

- No algorithm can obtain a regret bound better than $\Omega(\sqrt{T})$.
- Stochastic bandits are easier than adversarial bandits. Lower bounds for stochastic bandits are also applicable for adversarial bandits.
- Certifies the near-optimality of $\sqrt{T}$ regret for UCB [Auer et al., 2002a] and EXP3 [Auer et al., 2002b].

Instance-dependent lower bound

[Lai and Robbins, 1985]: we might be able to say something less pessimistic in an instance-dependent manner.

Theorem 4 (Instance-dependent lower bound)

Consider a strategy that satisfies $\mathbb{E}[R_T] = o(T^\alpha)$ *for any set of reward distributions* $\{\mathbb{P}_i\}_{1 \le i \le n}$ *indexed by a single real parameter, any arm* i *with sub-optimality gap* $\Delta_i > 0$, *and any* $\alpha > 0$. *Then, the following holds*

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \ge \sum_{i:\ \Delta_i > 0} \frac{\Delta_i}{\mathsf{KL}(\mathbb{P}_i \| \mathbb{P}^\star)},$$

where $\mathbb{P}^\star$ *is the distribution of the optimal arm.*

- The instance-dependent lower bound is $\Omega(\log T)$.

Near instance-optimality of UCB

For Gaussian bandits,

$$\mathsf{KL}(\mu_i \| \mu^\star) = \frac{\Delta_i^2}{2},$$

then it follows that

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq \sum_{i:\ \Delta_i > 0} \frac{2}{\Delta_i},$$

and

$$R_T \gtrsim \sum_{i:\ \Delta_i > 0} \frac{\log T}{\Delta_i}.$$

- This almost matches with the instance-dependent upper bound of UCB, which says (ignoring n)

$$R_T \lesssim \sum_{i:\Delta_i > 0} \frac{\log T}{\Delta_i}.$$

Analysis

KL Divergence between two bandits

Lemma 5 (Divergence decomposition)

- *Let $\nu = (\mathbb{P}_1, \ldots, \mathbb{P}_n)$ be the reward distributions associated with one n-armed bandit, and let $\nu' = (\mathbb{P}'_1, \ldots, \mathbb{P}'_n)$ be the reward distributions associated with another n-armed bandit.*
- *Fix some algorithm π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the bandit model $\{i_t, r_t\}_{t=1}^T$ induced by the T-round interconnection of π and ν (respectively, π and ν').*

Then,

$$\mathsf{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) = \sum_{i=1}^n \mathbb{E}_\nu[T_{i,T}] \mathsf{KL}(\mathbb{P}_i \| \mathbb{P}'_i),$$

where $T_{i,T} = \sum_{t=1}^T \mathbb{I}(i_t = i)$.

Bretagnolle-Huber inequality

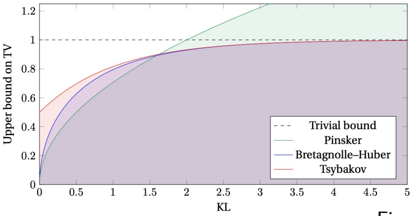


Figure credit: [Canonne, 2022].

Theorem 6 (Bretagnolle-Huber inequality)

For two probability distributions p, q over Ω, it follows that

$$d_{\mathsf{TV}}(p, q) \leq \sqrt{1 - e^{-\mathsf{KL}(p\|q)}} \leq 1 - \frac{1}{2}e^{-\mathsf{KL}(p\|q)}$$

- The second bound is due to [Tsybakov, 2008].
- As a consequence, for any event $A \subseteq \Omega$,

$$p(A^c) + q(A) \geq \frac{1}{2}e^{-\mathsf{KL}(p\|q)}.$$

Proof of minimax lower bound

Step 1: identifying a pair of bandits. Fix an algorithm π.

Suppose we begin with a Gaussian bandit ν with unit variance $\mathbb{P}_i = \mathcal{N}(\mu_i, 1)$, where $\mu^\star = \mu_1 > \mu_2 \geq \ldots \geq \mu_n$ w.l.o.g.. Let $\mathbb{P}_\nu$ be the resulting probability measure over T-round interconnection of π and ν.

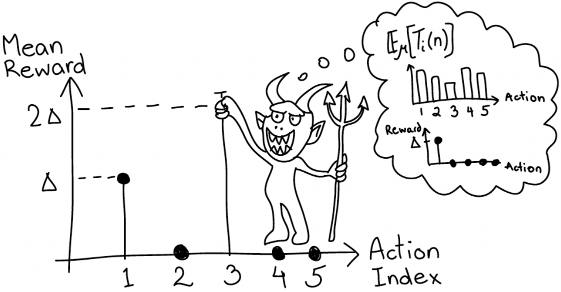


Figure 15.1 The idea of the minimax lower bound. Given a policy and one environment, the evil antagonist picks another environment so that the policy will suffer a large regret in at least one environment.

Figure credit: [Lattimore and Szepesvári, 2020]

Proof of minimax lower bound

Identifying the competing bandit: in view of the divergence decomposition lemma, let

$$j = \arg\min_{i>1} \mathbb{E}_{\nu}[T_{i,T}],$$

the arm that has been least pulled. The second bandit ν' is then selected as

$$\mathbb{P}'_i = \begin{cases} \mathbb{P}_i & i \neq j \\ \mathcal{N}(\mu_j + \lambda, 1) & i = j \end{cases},$$

where $\lambda > \Delta_j$ is to be selected. Arm j is optimal in the second bandit. Call the resulting probability measure $\mathbb{P}_{\nu'}$.

Proof of minimax lower bound

Step 2: computing the KL divergence. It is easy to observe that

$$\mathsf{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) = \mathbb{E}_\nu[T_{j,T}]\mathsf{KL}(\mathbb{P}_j \| \mathbb{P}'_j) \le \frac{T\lambda^2}{2(n-1)},$$

where we used

$$\sum_{i>1} \mathbb{E}_\nu[T_{i,T}] \le \sum_{i=1}^{n} \mathbb{E}_\nu[T_{i,T}] = T \qquad \Longrightarrow \qquad \mathbb{E}_\nu[T_{j,T}] \le \frac{T}{n-1}.$$

and

$$\mathsf{KL}\big(\mathcal{N}(\mu_j, 1) \| \mathcal{N}(\mu_j + \lambda, 1)\big) = \frac{\lambda^2}{2}.$$

Proof of minimax lower bound

Step 3: summing up the regrets of two bandits. By the regret decomposition lemma,

- For the first bandit ν, since j is sub-optimal,

$$R_T = \sum_{i \neq 1} \Delta_i \mathbb{E}_\nu[T_{i,T}] \geq \Delta_j \mathbb{E}_\nu[T_{j,T}] \geq \frac{T\Delta_j}{2} \mathbb{P}_\nu(T_{j,T} \geq T/2).$$

- For the second bandit, since j is optimal, for any $i \neq j$, it follows $\Delta'_i = \mu_j + \lambda - \mu_i = \lambda - (\mu_i - \mu_j) \geq \lambda - \Delta_j$, it follows

$$R'_T = \sum_{i \neq j} \Delta'_i \mathbb{E}_{\nu'}[T_{i,T}] \geq \frac{T(\lambda - \Delta_j)}{2} \mathbb{P}_{\nu'}(T_{j,T} < T/2).$$

Proof of minimax lower bound

Step 3: summing up the regrets of two bandits. By the regret decomposition lemma,

- For the first bandit ν, since j is sub-optimal,

$$R_T = \sum_{i \neq 1} \Delta_i \mathbb{E}_\nu[T_{i,T}] \geq \Delta_j \mathbb{E}_\nu[T_{j,T}] \geq \frac{T\Delta_j}{2} \mathbb{P}_\nu(T_{j,T} \geq T/2).$$

- For the second bandit, since j is optimal, for any $i \neq j$, it follows $\Delta_i' = \mu_j + \lambda - \mu_i = \lambda - (\mu_i - \mu_j) \geq \lambda - \Delta_j$, it follows

$$R_T' = \sum_{i \neq j} \Delta_i' \mathbb{E}_{\nu'}[T_{i,T}] \geq \frac{T(\lambda - \Delta_j)}{2} \mathbb{P}_{\nu'}(T_{j,T} < T/2).$$

Letting $A = \{T_{j,T} < T/2\}$, and summing these up, we have

$$R_T + R_T' \geq \frac{T}{2} \min\{\Delta_j, \lambda - \Delta_j\} \left[\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c)\right].$$

Proof of minimax lower bound

Step 4: finishing up by Bretagnolle-Huber. By Bretagnolle-Huber,

$$\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c) \geq \frac{1}{2} e^{-\mathsf{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'})} \geq \frac{1}{2} \exp\left(-\frac{2T\lambda^2}{(n-1)}\right).$$

$$\Longrightarrow \qquad R_T + R'_T \geq \frac{T}{4} \min\{\Delta_j, \lambda - \Delta_j\} \exp\left(-\frac{T\lambda^2}{2(n-1)}\right).$$

Setting $\lambda = 2\Delta_j$ leads to

$$R_T + R'_T \geq \frac{T}{4} \Delta_j \exp\left(-\frac{2T\Delta_j^2}{(n-1)}\right).$$

Let $\mu^\star = \mu_1 = \Delta$ and $\mu_2, \ldots, \mu_n = 0$. Set $\Delta_j = \Delta = \sqrt{(n-1)/4T} \leq 1/2$, we have

$$R_T + R'_T \geq \frac{e^{-1/2}}{8} \sqrt{(n-1)T} \Longrightarrow \max\{R_T, R'_T\} \geq \frac{e^{-1/2}}{16} \sqrt{(n-1)T}.$$

(Recall) instance-dependent lower bound

[Lai and Robbins, 1985]: we might be able to say something less pessimistic in an instance-dependent manner.

Theorem 7 (Instance-dependent lower bound)

Consider a strategy that satisfies $\mathbb{E}[R_T] = o(T^\alpha)$ for any set of reward distributions $\{\mathbb{P}_i\}_{1\le i\le n}$ indexed by a single real parameter, any arm i with sub-optimality gap $\Delta_i > 0$, and any $\alpha > 0$. Then, the following holds

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \ge \sum_{i:\ \Delta_i>0} \frac{\Delta_i}{\mathsf{KL}(\mathbb{P}_i \| \mathbb{P}^\star)},$$

where $\mathbb{P}^\star$ is the distribution of the optimal arm.

- The instance-dependent lower bound is $\Omega(\log T)$.

Proof of instance-dependent lower bound

We only consider **Gaussian bandits**.

In view of the regret decomposition lemma, it is sufficient to show for any sub-optimal arm i,

$$\text{liminf}_{T\to\infty} \frac{\mathbb{E}_\nu[T_{i,T}]}{\log T} \geq \frac{2}{\Delta_i^2}.$$

Let us fix a sub-optimal arm $j \neq i^\star$.

Step 1: identify the competing bandit. Motivated to the earlier proof, we construct second bandit ν' is then selected as

$$\mathbb{P}'_i = \begin{cases} \mathbb{P}_i & i \neq j \\ \mathcal{N}(\mu_j + \lambda, 1) & i = j \end{cases},$$

where we set $\lambda > \Delta_j$, and arm j is optimal in the second bandit.

Proof of instance-dependent lower bound

Step 2: lower bound the regret via Bretagnolle-Huber. Similar to the earlier proof, we obtain

$$\begin{aligned} R_T + R'_T &\geq \frac{T \min\{\Delta_j, \lambda - \Delta_j\}}{4} e^{-\mathsf{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'})} \\ &= \frac{T \min\{\Delta_j, \lambda - \Delta_j\}}{4} e^{-\lambda^2 \mathbb{E}_\nu [T_{j,T}]/2}, \end{aligned}$$

Proof of instance-dependent lower bound

Step 2: lower bound the regret via Bretagnolle-Huber. Similar to the earlier proof, we obtain

$$
\begin{aligned}
R_T + R'_T &\geq \frac{T \min\{\Delta_j, \lambda - \Delta_j\}}{4} e^{-\mathsf{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'})} \\
&= \frac{T \min\{\Delta_j, \lambda - \Delta_j\}}{4} e^{-\lambda^2 \mathbb{E}_\nu[T_{j,T}]/2},
\end{aligned}
$$

which gives

$$
\begin{aligned}
&\mathbb{E}_\nu[T_{j,T}] \geq \frac{2}{\lambda^2} \log\left(\frac{T \min\{\Delta_j, \lambda - \Delta_j\}}{4(R_T + R'_T)}\right) \\
\Longrightarrow \qquad &\frac{\lambda^2}{2} \frac{\mathbb{E}_\nu[T_{j,T}]}{\log T} \geq \left[1 + \frac{\log \min\{\Delta_j, \lambda - \Delta_j\}}{4 \log T} - \frac{\log(R_T + R'_T)}{\log T}\right].
\end{aligned}
$$

The next step is to examine the liminf of the right-hand-side.

Proof of instance-dependent lower bound

Step 3: taking limits to finish up. Since for any $\alpha > 0$, there exist some constant $C_\alpha > 0$ such that

$$R_T + R'_T \le C_\alpha T^\alpha$$

for all T, we have

$$\mathsf{limsup}_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} \le \mathsf{limsup}_{T\to\infty} \frac{\alpha \log T + \log C_\alpha}{\log T} = \alpha.$$

Proof of instance-dependent lower bound

Step 3: taking limits to finish up. Since for any $\alpha > 0$, there exist some constant $C_\alpha > 0$ such that

$$R_T + R'_T \leq C_\alpha T^\alpha$$

for all T, we have

$$\mathsf{limsup}_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} \leq \mathsf{limsup}_{T\to\infty} \frac{\alpha \log T + \log C_\alpha}{\log T} = \alpha.$$

Since this holds for any arbitrary $\alpha > 0$, it follows that

$$\mathsf{limsup}_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} = 0.$$

Proof of instance-dependent lower bound

Step 3: taking limits to finish up. Since for any $\alpha > 0$, there exist some constant $C_\alpha > 0$ such that

$$R_T + R'_T \leq C_\alpha T^\alpha$$

for all T, we have

$$\mathsf{limsup}_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} \leq \mathsf{limsup}_{T\to\infty} \frac{\alpha \log T + \log C_\alpha}{\log T} = \alpha.$$

Since this holds for any arbitrary $\alpha > 0$, it follows that

$$\mathsf{limsup}_{T\to\infty} \frac{\log(R_T + R'_T)}{\log T} = 0.$$

Consequently,

$$\mathsf{liminf}_{T\to\infty} \frac{\lambda^2}{2} \frac{\mathbb{E}_\nu[T_{j,T}]}{\log T} \geq 1.$$

Taking the infimum of both sides over $\lambda > \Delta_j$ thus finishes the proof.

Further pointers

The literature on bandits is vast, and we have only scratched the surface.

We will come back and visit some additional variations, e.g., when dealing with function approximation.

Further pointers to worthy topics:

- Thompson sampling: a Bayesian approach
- Beyond EXP3: dealing with variance

Excellent reference: Bandit algorithms [Lattimore and Szepesvári, 2020].

References I

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a).
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47(2):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b).
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77.

Canonne, C. L. (2022).
A short note on an inequality between KL and TV.
arXiv preprint arXiv:2202.07198.

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

References II

Tsybakov, A. B. (2008).
Introduction to Nonparametric Estimation.
Springer Publishing Company, Incorporated, 1st edition.